Density Based Clustering

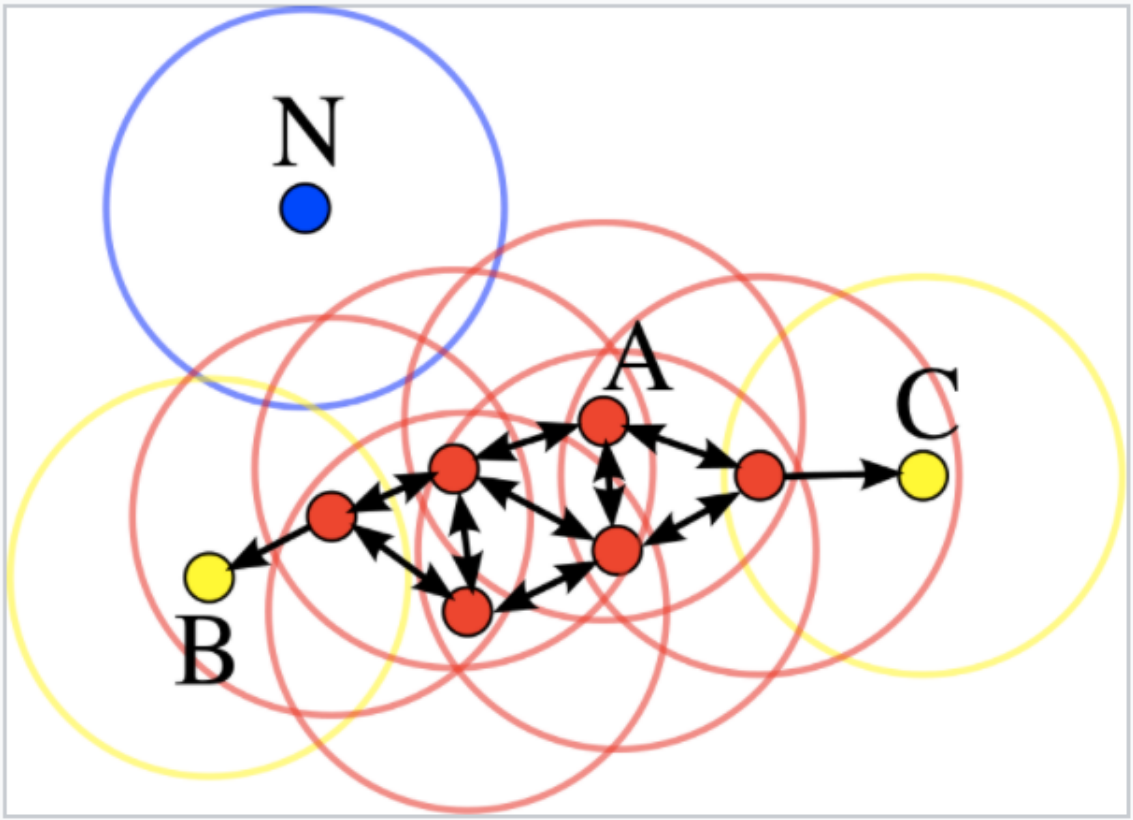
- Clustering based on density (local cluster criterion), such as density-connected points or based on an explicitly constructed density function
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan
  - Need density parameters

- Several interesting studies:
  - DBSCAN: Ester, et al. (KDD'96)
  - DENCLUE: Hinneburg & D. Keim (KDD'98/2006)
  - OPTICS: Ankerst, et al (SIGMOD'99).
  - CLIQUE: Agrawal, et al. (SIGMOD'98)

DBSCAN: Density-Based Algorithm for Discovering Clusters

- DBSCAN is a density-based algorithm.
  - Density = number of points within a specified radius r (Eps)
  - A point is a core point if it has more than a specified number of points (MinPts) within Eps
    - These are points that are at the interior of a cluster
  - A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point
  - A noise point is any point that is not a core point or a border point.

DBSCAN: Density-Based Algorithm for Discovering Clusters



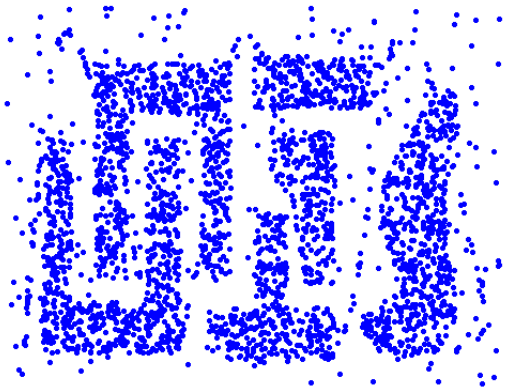
In this diagram, $\mathrm{minPts} = 4$. Point A and the other red points are core points, because the area surrounding these points in an ε radius contain at least 4 points (including the point itself). Because they are all reachable from one another, they form a single cluster. Points B and C are not core points, but are reachable from A (via other core points) and thus belong to the cluster as well. Point N is a noise point that is neither a core point nor directly-reachable.

Complexity

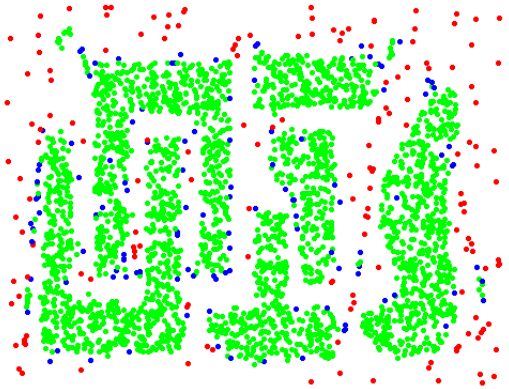
- Time Complexity: $O(n^2)$—for each point it has to be determined if it is a core point, can be reduced to O(n*log(n)) in lower dimensional spaces by using efficient data structures (n is the number of objects to be clustered);

- Space Complexity: O(n).

DBSCAN: Density-Based Algorithm for Discovering Clusters



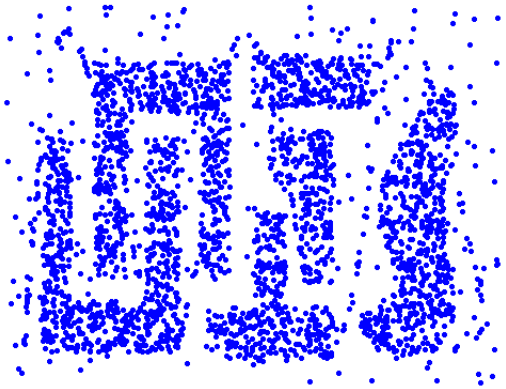
Original Points



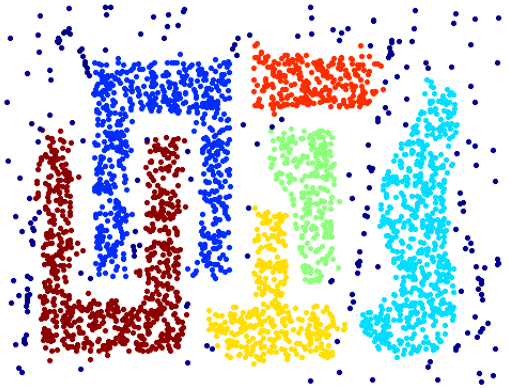
Point types: core, border and noise

Eps = 10, MinPts = 4

When DBSCAN works well?



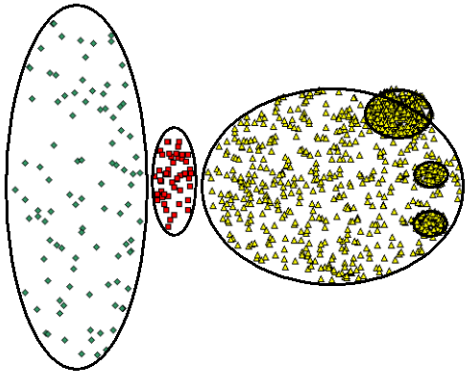
Original Points



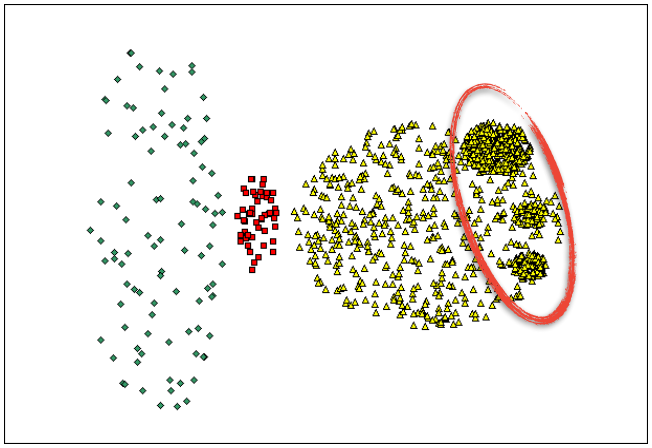
Clusters

- Resistant to Noise
- Can handle clusters of different shapes and sizes

When DBSCAN fails?

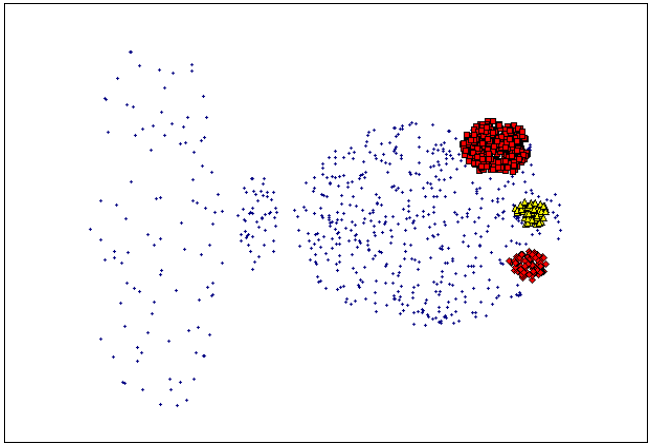


Original Points



Miss these clusters

(MinPts=4, Eps=9.75).



(MinPts=4, Eps=9.92)

- Varying densities
- High-dimensional data

Summery

- Good:
  - can detect arbitrary shapes,
  - not very sensitive to noise,
  - supports outlier detection,
  - complexity is kind of okay,
  - beside K-means the second most used clustering algorithm.

- Bad:
  - does not work well in high-dimensional datasets,
  - parameter selection is tricky,
  - has problems of identifying clusters of varying densities